The University Of Sheffield.

# High Level gene expression Analysis. Functional Annotation and Pathway Analysis

The University Of Sheffield.

# Today's Outline

Part A : Recap of the gene expression analysis workflow
When using different methods
Probabilistic models for gene expression data

*Break – 5 min (questions if any)*

Part B : Dimensionality reduction
Hierarchical clustering
Principal Component Analysis

Functional Annotation
Pathway analysis

Class activity: Discussion based on your questions from week 4 practical

The University Of Sheffield.

# Today we will be learning:

- How to build a complete workflow for gene expression analysis
- How to compare methods for estimation of gene expression on real research data
- High level analysis: Differential Expression
- Hierarchical clustering
- Principal Component Analysis
- Gene Ontologies
- Functional analysis
- Pathway analysis

The University Of Sheffield.

# High Level Analysis: workflow example

1. *Visualisation of the data*

2. *High level summary*: Combine expression from replicated arrays
   - Combine expression and uncertainty (puma)
   - Combine information from replicates with single poing statistics

3. *Differential Expression Analysis:* Determine differential expression between conditions, or between more complex contrasts such as interaction terms

4. *Dimensionality reduction* – Principal Component Analysis

5. *Data Clustering:* Cluster data taking the expression-level uncertainty into account

The University Of Sheffield.

# The Fold Change

Given two gene expression values **x** and **y** the fold change is defined as

$$FC = \frac{x}{y}$$

Given two vectors $\mathbf{X}_j$ and $\mathbf{Y}_j$ of gene expression measurements for controls and cases for GENE j, the fold change is defined as

$$FC_j = \frac{x_{ij}}{y_{ij}} \qquad i = 1, \ldots, n$$

where n is number of samples

It can also appear as a difference when we use the log transformation of the data.

Problem: How do we manage replicates? We need to combine the data

The University Of Sheffield.

# High Level summary

*High level summary*: Combine expression from replicated arrays

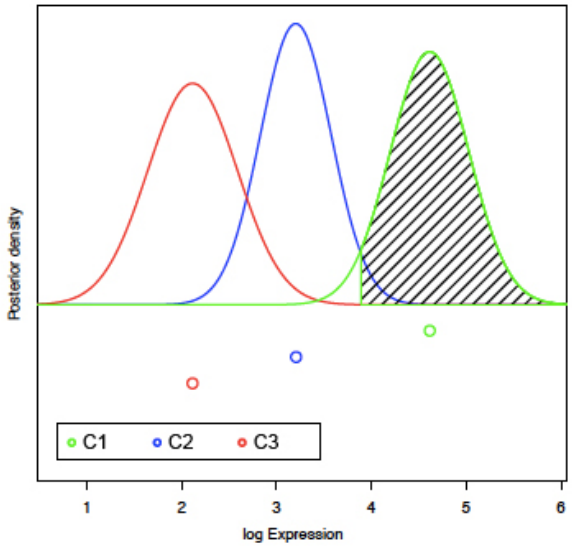
This is prior our DE and it is useful to have a representative gene expression value for the class you are studying : Wild Type vs Mutant; Disease vs Control.

- Combine information from replicates with single point statistics
- Combine expression and uncertainty (puma) using Bayesian Inference

In both cases you need a measure of uncertainty.

How do we get it in both cases?

# Differential Expression: pumaDE and PPLR

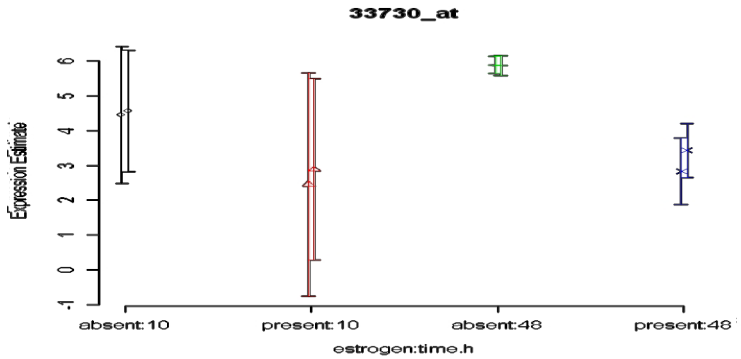



**Probability of Positive Log ratio: PPLR**

$$P(\mu_1 > \mu_2 | D, \phi) = \int_0^{\infty} P(\mu_1 - \mu_2 | D, \phi) d(\mu_1 - \mu_2)$$

**Example:**
In differential Expression analysis the goal is to estimate genes that change across conditions.

What happens then if we do not evaluate uncertainty?




Data from Choe et al, *Genome Biology*(2005)

The University Of Sheffield.

# Differential Expression: selecting the DE genes

**Single point statistics analysis:**

After combining the data we calculate :

FC
FDR and p-values (limma)
use a threshold |FC|>1 and set a FDR

**puma analysis:**

After combining the data with *pumaComb* we calculate :

FC
PPLR (pumaDE)
use a threshold PPLR>0.80 or higher for positively regulated
use a Threshold of PPLR<0.2 or lower for negatively regulated genes

MAKE SURE YOU LOOK AT THE OVERALL EXPRESSION

The University Of Sheffield.

# How do we use limma?

**Fist step :**
define the matrices

**Second step:**
Fit a linear model to explain the relation between each gene on the array and the design matrix

$$E(y_i) = X\alpha_i$$

Where $y_i$ is the gene expression for gene i, X is the design matrix and $\alpha_i$ is the coefficient for the gene i

**Third Step**: define significance levels for the comparisons. We use an empirical Bayes t-test, to calculate the p-values. This is to ensure that the information is learned from the data as a whole rather than single point.

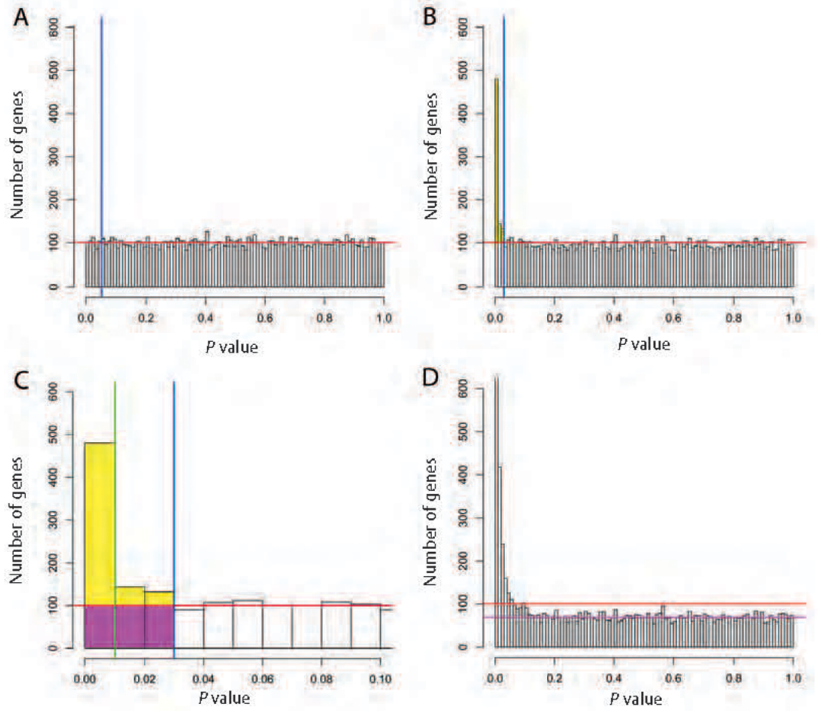
**Fourth Step**: Select the significant targets and visualise the data

# False Discovery Rate

10,000 genes

A. No DE
B. 500 DE
C. Zoom in
D. 1000 DE

TP:456
FP:300




when a significance level of 0.03 is used, the FDR = 300/756 = 0.40. Now we know that on average 40% of 756 genes are false positives

if we use a significance level of 0.01 (green line), the number of true positives is 380, and the number of false positives is 100, so the FDR = 0.21

At a significance level of 0.0007321 the FDR = 0.05, which corresponds to (on average) 141.5 true positives and 7.5 false positives – with Bonferroni an average of 3 true positives and no false positives.

The University Of Sheffield.

# Part B

The University Of Sheffield.

# How to discover patterns

There are three important questions we need to ask:

- Why do we want to discover patterns in gene expression data?
- What can we associate patterns to?
- What is our ultimate goal when studying biological systems?

The University Of Sheffield.

# Principal Component Analysis

It is one of the most commonly used technique to visualise and interpret high dimensional data

It identifies the maximum spread of the data maximising the variance by rotating the space where the data lives.

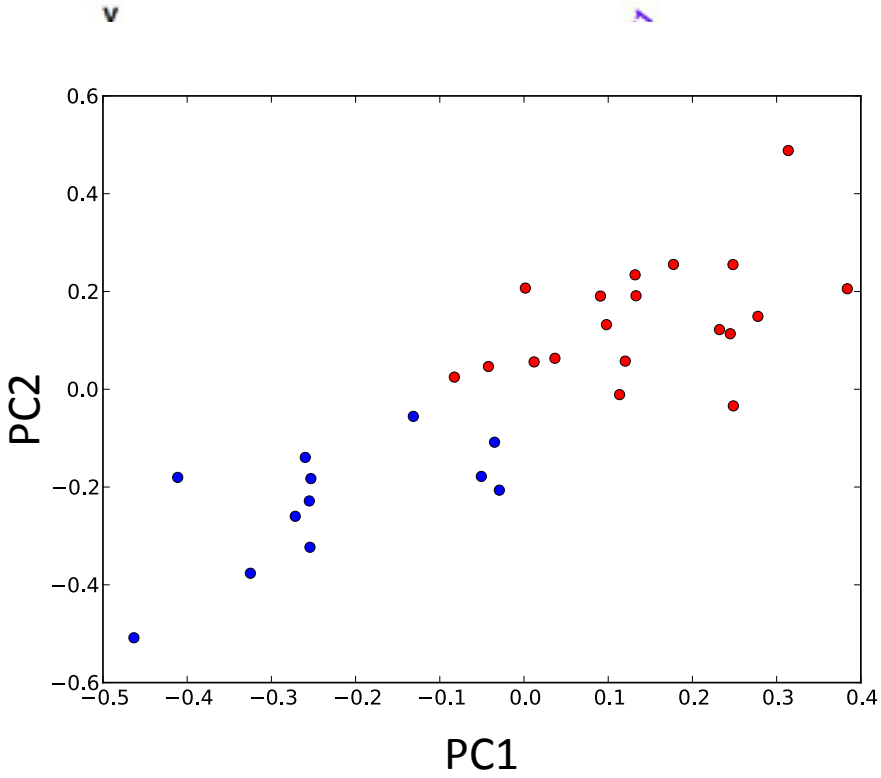
It uses a set of variables that are hidden to the user and are implicitly explained by the data (latent variables)

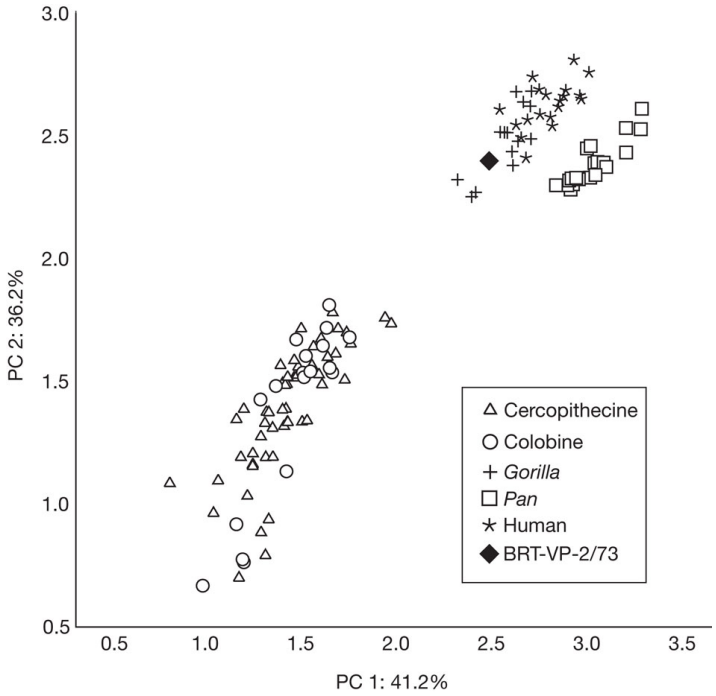
Every direction found that extract informative features from the “noisy” cloud of data points is called a principal component

Dimensionality reduction

The University Of Sheffield.

# Principal Component Analysis (*cont…)*




Y Haile-Selassie *et al*. *Nature* **483**, 565-569 (2012)
doi:10.1038/nature10922

usually reasonable, but it assumes that the uncertainty associated to each gene is constant

non-linear transformation of gene expression (Huber et al. 2002),PUMA PCA (Sanguinetti et al., 2005)

The University Of Sheffield.

# Clustering

- basic idea: group together genes that have similar pattern of expression across conditions or across time

- what do we mean by similar?

- different measures of similarity: Euclidean distance, angle
- between vectors, correlation coefficient, . . .

- Shared pattern of expression might be associate to similar functions

The University Of Sheffield.

# Similarity measures

A *similarity measure* (*function*) is a function that quantifies the similarity between two objects.

You can think of it as the inverse of distance metrics:

- large values for similar objects;
- zero or a negative value for very dissimilar objects.

E.g., in the context of cluster analysis we can use the following similarity measure:

$$d(X,Y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$ squared Euclidean distance

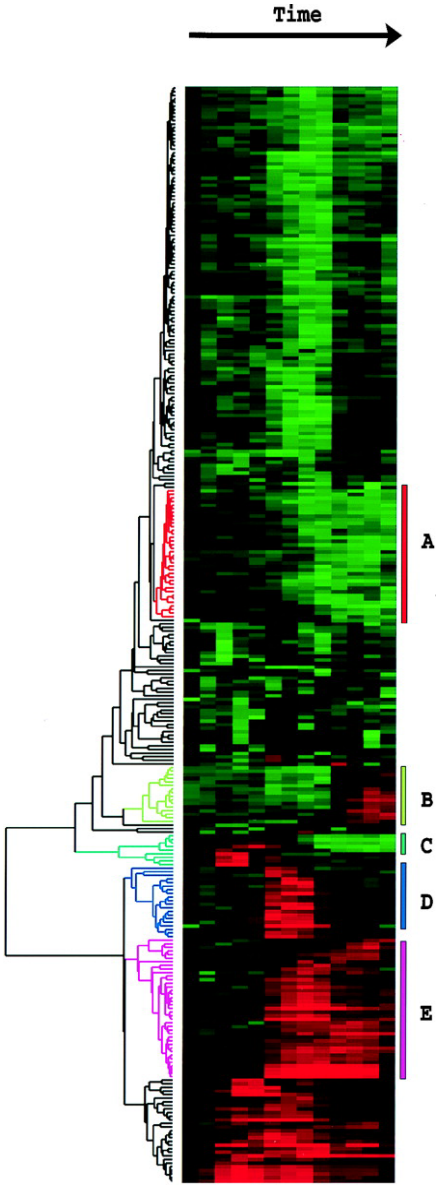
$$r_{XY} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$ Pearson Correlation

A similarity matrix is a matrix of scores that represent the similarity between a number of data points. Each element of the similarity matrix contains a measure of similarity between two of the data points. ( from *Wikipedia*)

The University Of Sheffield.

# Hierarchical Clustering

- builds a hierarchy of clusters
- bottom up (merging clusters) or top down (splitting clusters)
- Eisen et al. (1998). The genes that are most correlated are joined together, the expression value for the resulting node is the average expression of the two (or more) genes. The similarity matrix is then updated with the new node.
- Different similarity measures lead to different interpretations



The University Of Sheffield.

# Functional Analysis

The University Of Sheffield.

# Questions we will address

- What is functional analysis?
- Open source packages which enable to explore the pathways

Using methods like pathways analysis we can build:

- Gene Network – not in this module

The University Of Sheffield.

# Gene Ontologies

The **Gene Ontology (GO)** is a controlled vocabulary of terms to describe gene product characteristics in the domains of localization and function.

The aims of the GOs are:

- Maintain and develop its controlled vocabulary of gene and gene product attributes
- Annotate genes and gene products
- Acquire and disseminate annotation data
- Provide tools for easy access to the databases
- enable functional interpretation of experimental data using the GO

The ontologies covers these domains:

1. cellular component: the parts of a cell or its extracellular environment;
2. molecular function: the elemental activities of a gene product at the molecular level, for example as binding or catalysis;
3. biological process: operations or sets of molecular events

The Gene Ontology Project is what promoted and created this concept:

http://geneontology.org/

The University Of Sheffield.

# Pathway analysis

To identify interaction between genes based on literature knowledge: description of annotation and pattern of expression, association to disease etc.

There are two ways of approaching this type of analysis:

**Top down or bottom up**

*Top down*:
Look at the whole organism and abstract large portions of it

*Bottom up*:
Try to understand each small piece and assemble into the whole

Both are used, valid and interconnect.

The University Of Sheffield.

# Pathway analysis (*cont…*)

Biological annotations have started to include descriptions of gene interactions in the form of gene signaling networks, such as KEGG (Ogata et al.,1999), BioCarta (www.biocarta.com) this makes the pathway analysis *in silico* more informative than it was in the past.

There is a variety of open source software available for this.
There are also some very good commercial packages

We will explore the two main pathways navigators based on annotations and protein interactions.

The University Of Sheffield.

# DAVID

The Database for Annotation, Visualization and Integrated Discovery (**DAVID** )

http://david.abcc.ncifcrf.gov/

Entirely based on ontologies and annotations relies on KEGG and BioCarta

The University Of Sheffield.

# PANTHER

The **PANTHER** (Protein ANalysis THrough Evolutionary Relationships)
It is a Classification System designed to classify proteins (and their genes)

Proteins have been classified according to:
Family and subfamily: families are groups of evolutionarily related proteins; subfamilies are related proteins that also have the same function

Molecular function: the function of the protein by itself or with directly interacting proteins at a biochemical level, e.g. a protein kinase

Biological process: the function of the protein in the context of a larger network of proteins that interact to accomplish a process at the level of the cell or organism, e.g. mitosis.

Pathway: similar to biological process, but a pathway also explicitly specifies the relationships between the interacting molecules.

http://www.pantherdb.org/about.jsp

The University Of Sheffield.

# What else …

PANTHER, to define a list of enriched pathways that are represented by the list of genes that you are interested in.

You can also look at more systems biology oriented approaches, where probabilistic models are used to manage the annotations and the gene interactions, but this is not part of the module

The University Of Sheffield.

# Summary

A complete workflow for gene expression data analysis comprises:

- Low level analysis
- Diagnostics
- High level analysis
- Functional analysis
- Pathways analysis

Annotation of our targets using GOs is important for the biological interpretation of the data